Assignment 2 - CSC/DSC 265/465 - Spring 2019 - Due February 28

Unless otherwise specified, statistical significance can be accepted when the relevant P-value is no larger than $\alpha = 0.05$. Note that problem **Q5** is reserved for graduate students.

Q1: Recall the *negative binomial* random variable $X \sim nb(r, p)$. Suppose $U_1, U_2, \ldots$ is an infinite sequence of independent Bernoulli random variables with parameter p. Then let X be the number of such random variables required to observe r 1's. Then X has probability mass function (PMF)

$$p_X(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad x = r, r+1, \ldots.$$

It can then be shown that the mean of X is $E[X] = \mu = r/p$.

(a) Show that given single observation $X \sim nb(r, p)$, the maximum likelihood estimate of p is $\hat{p}_{MLE} = r/X$, and that, therefore, the maximum likelihood estimate of μ is

$$\hat{\mu}_{MLE} = \frac{r}{\hat{p}_{MLE}} = X.$$

(b) Suppose in the context of Bayesian inference, we fix r, and assign a $beta(\alpha, \beta)$ prior distribution to p (see Section 8.2.2 of the lecture notes). We then observe a single negative binomial random variable $X \sim nb(r, p)$. Show that the beta density is a conjugate prior for p. Give the exact posterior density of p, given observation $X = x$.

(c) Assuming $\alpha > 1$, what is the expected value of $\mu = r/p$ under the prior distribution of p (say $\hat{\mu}_{prior}$) and under the posterior distribution of p given observation $X = x$ (say $\hat{\mu}_{post}$)? **HINT:** Following the analysis given in Section 8.2.2, what is $E[Z^{-1}]$ if $Z \sim beta(a, b)$?

(d) Show that we can write

$$\hat{\mu}_{post} = q\hat{\mu}_{MLE} + (1-q)\hat{\mu}_{prior}$$

where q depends only on α and r.

Q2: Suppose we observe survival times $105, 107.5, 107.5, 107.5_+, 110_+, 115, 115, 120$, where T_+ is a right-censored survival time.

(a) Consider time points $(t_0, t_1, t_2, t_3, t_4, t_5) = (0, 105, 107.5, 110, 115, 120)$. For each of these times t' give the number *at risk* at time t' (the number of subjects with survival times $T \geq t'$ or $T_+ \geq t'$), and the number who die at time t' ($T = t'$). Summarize these quantities in a tabular form.

(b) Plot a Kaplan-Meier estimate of the survival curve. Do this two ways:

(i) Plot the curve directly from the numbers in the table. Use the `pty='s'` option. Indicate all points on the curve at which an observation was censored (use a `+` symbol (`pch=3`)).

(ii) Use the `survfit()` function. Set options `conf.int=FALSE` and `mark.time=TRUE`.

Are the two curves the same?

Q3: Load the data frame `VA` from the library `MASS`. This data set is from a Veteran's Administration lung cancer trial. It includes variables `stime` (survival or follow-up time in days since initial observation), `status` (dead (=1) or censored (=0) at time `stime`), `cell` (factor indicating one of 4 cell types), `Karn` (Karnofsky score of patient's performance on a scale of 0 to 100). If `status == 1` the patient has died at time `stime`, otherwise the survival time given in `stime` is right-censored.

The Cox proportional hazards model predicts the hazard rate

$$h(x) = h_0(x)e^{\eta}, \quad x > 0,$$

where any predictor is incorporated into η, but not the baseline hazard rate $h_0(x)$. Therefore, a crucial assumption is that the hazard rate functions of all observations are proportional to h_0, and therefore to each other. In this example, this can be checked by comparing the hazard rates estimated separately for each of the four cell types.

(a) Show that if the hazard rates are proportional, the cumulative hazard rates will be proportional as well.

(b) Calculate Kaplan-Meier estimates of the survival curves separately for each cell type. Plot the cumulative hazard functions on a single graph. Does the proportional hazards assumption seem reasonable? **HINT:** See software `SURVIVAL.R` available on BLACKBOARD for sample code.

(c) Fit three Cox proportional hazards models using the `coxph` function:

(A) With factor variable `cell` as the sole predictor.

(B) With `cell` and `Karn` as additive predictors.

(C) With `cell` and `Karn`, including all interactions.

Is model (B) a significant improvement over model (A)? Is model (C) a significant improvement over model (B)? **HINT:** You can use the `anova` method to compare nested models, in this case based on a χ^2 test.

(d) Given a Cox proportional hazards model fit using the `coxph` function, the `predict` method can be used to obtain estimates of the survival curve at any time t for any combination of covariate values. For example, suppose we fit a Cox proportional hazards model

```
fit.cox = coxph(Surv(time,status) ~ group + age)
```

where `group` is a factor variable with levels `'male', 'female'` and `age` is a quantitive age variable (in years, say). To estimate a survival curve for male subjects of age 45, we would define a time grid on which to evaluate the survival curve, then create a new data frame containing this time grid, as well as the desired covariate values:

```
time.grid = seq(0,100,1)
new.x = data.frame(time = time.grid, status = rep(1,101), group = rep('male',101),
                age = rep(45,101))
pred.curve = predict(fit.cox,newdata=new.x,type='expected')
plot(time.grid,exp(-pred.curve))
```

For each cell type $k = 1, 2, 3, 4$, create a separate plot containing

- A Kaplan-Meier survival curve estimate using only data for cell type k.
- An estimated survival curve based on the Cox proportional hazards model for cell type k, and `Karn` $= Q_1$, where Q_1 is the 25th percentile of the available Karnofsky scores.
- An estimated survival curve based on the Cox proportional hazards model for cell type k, and `Karn` $= Q_3$, where Q_3 is the 75th percentile of the available Karnofsky scores.

Do this using the additive model (B), then again with the interactive model (C) (there should be a total of eight separate plots). Use the `legend` function to clearly label each survival curve estimate. **HINT:** The `subset` option can be used in many model fitting functions to select a subset of the data to be used for the fit.

(e) If the interactive model (C) is a significant improvement over the additive model (B), this can be interpreted to mean that the relationship between a Karnofsky score and a patients survival time differs by cell type. Based on Parts (c) and (d), can you conclude that the predictors `cell` and `Karn` are interactive in this way?

Q4: One important application in remote sensor systems is the localization of a system node based on *received signal strength indication* (RSSI) measurements. Suppose m stationary nodes equipped with radio

signal receivers are deployed in an environment, which is mathematically a region within $\mathbb{R}^2$. Suppose an additional mobile node to be localized is equipped with a radio signal transmitter. The strength of the signal (RSSI) received at each stationary node is inversely proportional to the distance between that node and the mobile node (that is, the transmission distance).

The *Weibull density* is often used to model survival times, and other positive random variables. Parameterizations vary, but it commonly incorporates a *shape parameter* $\kappa > 0$ and *scale parameter* $\mu > 0$ as follows:

$$f(z;\kappa,\mu) = \frac{\kappa}{\mu}\left(\frac{z}{\mu}\right)^{\kappa-1} e^{-(z/\mu)^\kappa}, \;\; z \geq 0.$$

Then suppose Y is an RSSI measurement which is calibrated so that $Z = 1/Y$ has a Weibull distribution where $\kappa = 8.25$ and μ is the transmission distance. There are $m = 3$ stationary nodes, labeled $i = 1, 2, 3$, located on the two-dimensional deployment region at coordinates (in miles):

$$\begin{aligned}(x_1, y_1) &= (-0.50, 0.00)\\ (x_2, y_2) &= (0.42, 2.00)\\ (x_3, y_3) &= (1.50, 1.27).\end{aligned}$$

Suppose $\theta = (\theta_x, \theta_y)$ is the current location of the mobile node. After transmitting a radio signal, three RSSI measurements $(Y_1, Y_2, Y_3) = (0.926, 0.943, 0.787)$ are collected from the respective stationary nodes. Assume the RSSI measurements are independent.

(a) The object is to construct a Bayesian model for the inference of $\theta = (\theta_x, \theta_y)$. For the prior density $\pi(\theta)$, use a uniform distribution over some large enough region containing all nodes. Then accept as the data the reciprocals $Z = (Z_1, Z_2, Z_3) = (1/Y_1, 1/Y_2, 1/Y_3)$. Write explicitly the conditional density of Z given θ, say $f(z_1, z_2, z_3 \mid \theta)$, and then give the posterior density of θ given $Z = z$, say $\pi(\theta \mid z_1, z_2, z_3)$. Note that the posterior density need not be normalized.

(b) Evaluate the posterior density on a two-dimensional grid, with x and y coordinates defined as follows

```
xgrid = seq(-0.7,1.0,0.01)
ygrid = seq(-0.1,1.8,0.01)
```

That is, $\pi(\theta \mid z_1, z_2, z_3)$ is evaluated in a rectangle $\theta_x \in [-0.7, 1.0]$ and $\theta_y \in [-0.1, 1.8]$, with spacing $\delta = 0.01$ between grid points in each axis direction. Use the data given here for the values (z_1, z_2, z_3). Use the `persp` function to create a 3-dimensional image of this density. Use options `theta=0, phi=30`. **HINT:** In `R`, densities and distribution functions can usually be evaluated on a logarithmic scale, using the `log = TRUE` or `log.p = TRUE` option, as appropriate. This is far preferable for an application like this. A good strategy would be to write a function which evaluates and returns the density on a logarithmic scale, which can later be exponentiated if needed. Again, note that the posterior density need not be normalized.

(c) In this example, the maximum likelihood estimate $\hat{\theta}_{MLE}$ will be the value of θ which maximizes $\pi(\theta \mid z_1, z_2, z_3)$ (assuming there is only one global maximum). Why is this the case? Use the grid evaluation of Part (b) to obtain an approximation of $\hat{\theta}_{MLE}$. **HINT:** This can be done with the `which` function, using the `arr.ind = TRUE` option.

(d) Create a Hastings-Metropolis algorithm to simulate a sample from $\pi(\theta \mid z_1, z_2, z_3)$. Implement the following features:

(i) Use as a proposal rule something like `theta.new = theta.old + runif(2,-1/10,1/10)`. This means the resulting state space is not discrete, but the algorithm will work in much the same way. Under this proposal rule we can take $1 = Q(\theta_2 \mid \theta_1)/Q(\theta_1 \mid \theta_2)$ when calculating the acceptance probability.

(ii) Allow $N = 100,000$ transitions. Capture in a single object all sampled values of θ. You can used $\theta = (0, 0)$ as the initial state. **HINT:** When constructing an MCMC algorithm, rather than calculate a ratio of densities, it is better to calculate a difference Δ in log-densities, and then calculate the exponential function of the difference, that is, e^Δ. For this reason, the function constructed in Part (b) should return log-densities.

(e) The posterior density is defined on a two-dimensional plane. There are a number of `R` functions that can be used to visualize functions or densities defined on $\mathbb{R}^2$.

(i) Use the `smoothScatter` function to plot the sampled θ values (this function is part of the `graphics` library). Rather than draw a simple scatter plot, this function draws a heat map representation of the sample density. Essentially, a deeper color shade indicates a higher density of sampled points. Use options `xlim=c(-1,1.5),ylim=c(-0.5,2.1)`. Include horizontal and vertical axes lines which pass through the origin (0,0) (you can use the `abline` function for this). Also indicate the locations of the three stationary nodes (you can use the `points` function for this). Then indicate the location of $\hat{\theta}_{MLE}$, using a distinct symbol for clarity.

(ii) Superimpose on your plot a contour plot of the posterior density evaluated in Part (b). Use the `contour` function with the `add = TRUE` option.

(f) One advantage of the use of MCMC sampling to estimate posterior densities is that various forms of inference become easy to implement. For example, suppose there is interest in the distance of the mobile node from the origin $(0, 0)$. Use the MCMC to estimate the posterior density of this distance. This can be done by first transforming the sampled values of θ, then constructing a histogram of these transformed values.

Q5: [For Graduate Students] Note that throughout this problem you may assume that the order of differentiation and integration is exchangeable where needed. However, this holds only under certain regularity conditions. See Casella and Berger, *Statistical Inference (1st or 2nd ed)*, for a comprehensive discussion of this issue in the context of probability and statistical theory.

(a) Suppose X is a random variable. The *moment generating function* is a function of a real variable t, defined by

$$M_X(t) = E\left[e^{tX}\right]$$

for any fixed t. Assuming that $M_X(t)$ is finite in some open interval (a, b), where $a < 0$ and $b > 0$, show that the kth moment of X can be calculated by the kth derivative of $M_X(t)$ evaluated at $t = 0$, that is,

$$\left.\frac{d^k M_X(t)}{dt^k}\right|_{t=0} = E[X^k], \;\; k = 1, 2, \ldots.$$

(b) Suppose a random variable X possesses a density from a parametric family $f(x \mid \theta)$, where θ is a parameter from a parameter space $\theta \in \Theta$. This parametric family is called a one dimensional *exponential family* if it can be written

$$f(x \mid \theta) = \exp\{\eta(\theta)T(x) - B(\theta)\}\, h(x),$$

for some functions $\eta(\theta), T(x), B(\theta), h(x)$, with $\Theta \subset \mathbb{R}$. An exponential family is expressed in its *natural parameterization* if it can be written:

$$f(x \mid \eta) = \exp\{\eta T(x) - A(\eta)\}\, h(x),$$

so that η is the *natural parameter*. Many important parametric families are one-dimensional exponential families (Poisson, geometric, binomial), while the uniform parametric family $f(x \mid \theta) = \theta^{-1} I\{0 < x < \theta\}$, $\theta \in (0, \infty)$, is not. The normal distribution $N(\mu, \sigma^2)$ is a two-dimensional exponential family with parameter $\theta = (\mu, \sigma^2)$.

Prove that the MGF of $T(X)$ is $m(t) = \exp\{A(\eta + t) - A(\eta)\}$, from which it follows that the mean and variance of $T(X)$ are $A'(\eta)$, $A''(\eta)$, the first and second derivatives of A evaluated at η.

(c) Prove that if $\hat{\eta} = \hat{\eta}(X)$ is any solution to the equation

$$T(X) = E_\eta\left[T(X)\right]$$

wrt η, then it uniquely maximizes the log-likelihood function $\ell(\eta; X)$. **HINT:** Show that the log-likelihood function of the natural parameter, $L(\eta; X)$, is strictly concave.